Bayesian Markov Blanket Estimation

Dinu Kaufmann[1], **Sonali Parbhoo, Aleksander Wieczorek,**
Sebastian Keller, David Adametz, Volker Roth
Departement of Mathematics and Computer Science, University of Basel, Switzerland

Abstract

This paper considers a Bayesian view for estimating the Markov blanket of a set of query variables, where the set of potential neighbours is big. We factorise the posterior such that the Markov blanket is conditionally independent of the network of the potential neighbours. By exploiting this blockwise decoupling, we derive analytic expressions for posterior conditionals. Subsequently, we develop an inference scheme, which makes use of the factorisation. As a result, estimation of a sub-network is possible without inferring an entire network. Since the resulting Gibbs sampler scales linearly with the number of variables, it can handle relatively large neighbourhoods. The proposed scheme results in faster convergence and superior mixing of the Markov chain than existing Bayesian network estimation techniques.

1 INTRODUCTION AND RELATED WORK

Estimating a network of dependencies among a set of objects is a difficult problem in statistics, particularly in high-dimensional settings or where the observed measurements are noisy. Gaussian Graphical Models (GGM) are a tool for representing such relationships in an interpretable way. In a classical GGM setting, the sparsity pattern of the inverse covariance matrix $\mathbf{W} = \mathbf{\Sigma}^{-1}$ encodes conditional independence between variables of the graph. Consequently, various estimators have been proposed that reduce the number of parameters by imposing sparsity constraints on $\mathbf{W}$. Among these, the popular graphical lasso procedure [Friedman et al., 2008; Meinshausen and Bühlmann, 2006] places a Wishart likelihood on the sample covariance and computes a point estimate of the graph by minimizing the penalised log-likelihood.

In situations where the variables in the network have different types, it is often more interesting to examine the connections between these types as opposed to estimating an entire network of all the associations. Consider the example in gene analysis where the dependency between only a few clinical factors and thousands of genetic markers is required. When we would like to focus on a particular portion of the network, it is useful to limit the estimate to the *Markov blanket* of the nodes we are interested in. These are the set of nodes that, when conditioned on, render the nodes of interest conditionally independent of the rest of the network.

In this paper we provide a Bayesian perspective of estimating the Markov blanket of a set of p query variables in an undirected network. Unlike the point estimate of the graphical lasso, the Bayesian view enables the computation of a posterior distribution of the Markov blanket. A Bayesian interpretation of the graphical lasso is presented by Wang et al. [2012]. This approach partitions the matrix $\mathbf{W}$ as shown on the left in Fig. 1. Posterior inference involves iterating through the individual variables to estimate the entire network. In particular, inference of the $\mathbf{W}_{12}$ block relies on estimating both $\mathbf{W}_{11}$ and $\mathbf{W}_{22}$. However, the coupling of $\mathbf{W}_{12}$ and $\mathbf{W}_{22}$ is a limiting factor that can be avoided in the context of Markov blanket estimation. This idea forms the basis of our paper. An important observation for the model we present here, is that the Wishart likelihood may be factorised such that the blocks $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ are de-coupled from $\mathbf{W}_{22}$. This result is provided as Lemma 1 in Section 2. We show that by combining the factorised likelihood with an appropriate choice of prior, we obtain a posterior distribution that preserves

[1]Contact e-mail addresses: {dinu.kaufmann, sonali.parbhoo, aleksander.wieczorek, sebastianmathias.keller, david.adametz, volker.roth} @unibas.ch

Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 51.

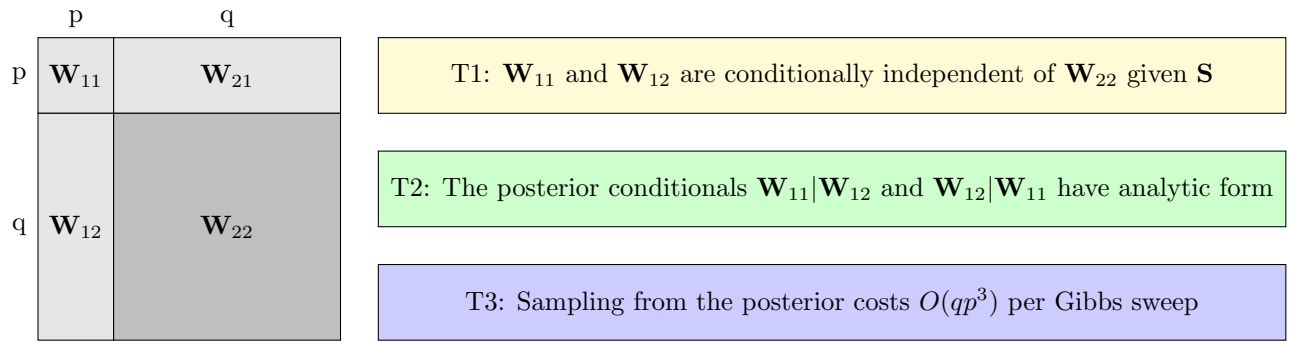

Figure 1: Overview of Bayesian Markov blanket estimation and key results.

this independence structure. Most importantly, this posterior distribution has an analytic form and can hence be sampled from. We formalise this in Section 3 as Theorem 2. A further consequence of this result is Theorem 3 which demonstrates that sampling from the posterior distribution can be done efficiently. Overall, this means that the Markov blanket of p query nodes, can be estimated efficiently *without explicitly inferring the entire network*.

An overview of our approach is presented in Fig. 1, where the matrix $\mathbf{W}$ is partitioned similarly to Wang et al. [2012]. More precisely, we consider the case where $p > 1$. The difference in the shading of the blocks in $\mathbf{W}$ indicates that estimation of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$ (and hence $\mathbf{W}_{21}$) is invariant of estimating $\mathbf{W}_{22}$.

The remainder of this paper is structured as follows. We begin by exploring the block factorization of the Wishart likelihood in Section 2. We subsequently derive the posterior distribution and construct a Gibbs sampler to efficiently sample from the different blocks in Section 3. Section 4 describes how Bayesian Markov blanket estimation can be extended to deal with mixed data types with the copula framework. Finally, we demonstrate the practical applicability of the scheme in Section 5 with examples of artificial and real data.

2 MODEL

Problem Formulation Assume $\mathbf{X} \in \mathbb{R}^{(p+q)\times n}$ is a given matrix with n independent observations. We are interested in estimating the Markov blanket of p query variables with respect to the q remaining variables in the data matrix. The sample covariance $\mathbf{S} = \mathbf{X}^\intercal\mathbf{X}$ follows the Wishart distribution $\mathbf{S} \sim \mathcal{W}_{p+q}(n, \mathbf{\Sigma})$ with n degrees of freedom. That is, $p(\mathbf{S}) \propto \det \mathbf{W}^{\frac{n}{2}} \exp \operatorname{tr}\left(-\frac{1}{2}\mathbf{WS}\right)$. Assume that $\mathbf{S}$ and $\mathbf{W}$ are partitioned according to

$$\mathbf{S} = \begin{array}{c} \\ p \\ q \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{pmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{pmatrix} \end{array}, \quad \mathbf{W} = \begin{array}{c} \\ p \\ q \end{array}\!\!\begin{array}{c} \begin{array}{cc} p & q \end{array} \\ \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix} \end{array},$$

where the matrices have been reordered such that the query variables lie in the upper left block. Given $\mathbf{S}$, we would like to infer $\mathbf{W}_{12}$, the Markov blanket of the p variables that constitute the block $\mathbf{S}_{11}$. We restrict the problem to the case where $p \ll q$ such that $\mathbf{S}_{11}$ is small, corresponding to the few variables of interest, and $\mathbf{S}_{22}$ is large.

Factorising the Likelihood We begin by showing a blockwise factorisation of the likelihood, which builds the foundation of our model. Let $\mathbf{W}_{22.1} = \mathbf{W}_{22} - \mathbf{W}_{21}\mathbf{W}_{11}^{-1}\mathbf{W}_{12}$ be the Schur complement of the block $\mathbf{W}_{11}$ in $\mathbf{W}$.

Lemma 1. *The likelihood of the covariance matrix factorises in terms of* $\mathbf{W}$ *as follows:*

$$\mathcal{L}_{\mathbf{S}}(\mathbf{W}) \propto \mathcal{L}_1(\mathbf{W}_{11}, \mathbf{W}_{12})\mathcal{L}_2(\mathbf{W}_{22\cdot 1}).$$

The proof of this lemma can be found in the supplementary document, and is analogous to Gupta and Nagar [1999] (Chapter 3, pp. 94–95). Lemma 1 is a pure functional statement without any statistical reasoning. The factorisation of the likelihood in Lemma 1 then translates to the analogous independence structure in the posterior distribution of $\mathbf{W}$ as shown in Theorem 1.

2.1 Prior

The natural conjugate prior to our likelihood is the Wishart distribution. However, in order to ensure sparsity, we also use a double exponential prior as in Wang et al. [2012]. Since our focus is on the Markov blanket, we only place the latter on the block $\mathbf{W}_{12}$.

This results in a compound prior:

$$\begin{aligned} p(\mathbf{W}) &= \mathcal{W}(p+q+1,\mathbf{I})p\left(\mathbf{W}_{12}|\mathbf{T}\right)p\left(\mathbf{T}|\gamma\right) \\ &\propto \exp\mathrm{tr}\left(-\tfrac{1}{2}\mathbf{I}\mathbf{W}\right) \\ &\prod_{w_{ij}\in\mathbf{W}_{12}} \tfrac{1}{\sqrt{2\pi t_{ij}}}\exp\left(-\frac{w_{ij}^2}{2t_{ij}}\right)\tfrac{\gamma^2}{2}\exp\left(-\tfrac{\gamma^2}{2}t_{ij}\right), \end{aligned} \tag{1}$$

where $\mathbf{T}=\{t_{ij}\}$ are inverse-Gaussian distributed scale parameters introduced by Wang et al. [2012]:

$$t_{ij}^{-1} \sim \mathcal{IG}\left(\sqrt{\gamma^2/w_{ij}^2},\gamma^2\right) \tag{2}$$

and γ is a hyperparameter. Most importantly, the compound prior also possesses the factorisation in terms of $\mathbf{W}$ proved for the likelihood in Lemma 1. This follows from from the element-wise independence of the prior. Multiplying the compound prior introduced in Eq. (1) by the likelihood yields the posterior distributions for blocks $\mathbf{W}_{12}$ and $\mathbf{W}_{11}$.

2.2 Posterior Distribution

A consequence of the factorisation in Lemma 1 is that the posterior distributions of the blocks $(\mathbf{W}_{11},\mathbf{W}_{12})$ and $\mathbf{W}_{22\cdot1}$ are conditionally independent given $\mathbf{S}$.

Theorem 1. *The posterior distribution of* $(\mathbf{W}_{11},\mathbf{W}_{12})$ *is conditionally independent of* $\mathbf{W}_{22\cdot1}$ *given* $\mathbf{S}$.

Proof. The Likelihood, as shown in Lemma 1, as well as the element-wise independent prior in Eqs. (1) and (2) factorise according to the blocks $\mathbf{W}_{11}$, $\mathbf{W}_{12}$, and $\mathbf{W}_{22}$. □

Because of the conditional independence proved in Theorem 1, we can infer the Markov blanket $\mathbf{W}_{12}$ without the need of estimating the big block $\mathbf{W}_{22\cdot1}$. In the next section, we explicitly derive the posterior distribution and show that it has an analytical form.

3 POSTERIOR INFERENCE

We now state the main result of this paper. Specifically, we show that the posterior distribution required to estimate the Markov blanket can be expressed in an analytical form. Subsequently, we demonstrate how to efficiently sample from it in Section 3.1.

Let the Matrix Generalised Inverse Gaussian (MGIG) distribution [Butler, 1998] be defined by probability density function with parameter λ:

$$\begin{aligned} &p(\mathbf{M};\lambda,\mathbf{A},\mathbf{B}) \\ &\quad\propto \det(\mathbf{M})^{-\lambda-1}\exp\mathrm{tr}\left(-\frac{1}{2}(\mathbf{A}\mathbf{M}+\mathbf{B}\mathbf{M}^{-1})\right). \end{aligned} \tag{3}$$

Theorem 2. *The posterior conditionals* $\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{S},\mathbf{T}$ *and* $\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{S},\mathbf{T}$ *admit an analytical form:*

(1) *Vectorised rows of* $\mathbf{W}_{12}$ *follow a joint normal distribution*

$$\begin{aligned} &vec(\mathbf{W}_{12}^{\mathsf{T}})|\mathbf{W}_{11},\mathbf{S},\mathbf{T} \\ &\sim \mathcal{N}_{pq}\left(vec\left(-(\mathbf{S}_{22}+\mathbf{I})^{-\mathsf{T}}\mathbf{S}_{12}^{\mathsf{T}}\mathbf{W}_{11}^{\mathsf{T}}\right),\mathbf{C}^{-1}\right), \end{aligned} \tag{4}$$

where $\mathbf{C}=\mathbf{W}_{11}^{-1}\otimes(\mathbf{S}_{22}+\mathbf{I})+\mathrm{diag}\left(\mathbf{D}_1,\ldots,\mathbf{D}_p\right)$ *be the covariance matrix, and* $\mathbf{D}_i=\mathrm{diag}\left((T_{i\cdot})^{-1}\right)$ *be diagonal matrices containing* $T_{i\cdot}=(t_{i1},\ldots,t_{iq})$.

(2) $\mathbf{W}_{11}$ *follows the Matrix Generalised Inverse Gaussian (MGIG) distribution:*

$$\begin{aligned} \mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{S},\mathbf{T} \sim \mathcal{MGIG}_{p\times p}\Big(&\frac{n}{2}+p, \\ &\mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{21},\mathbf{S}_{11}+\mathbf{I}\Big). \end{aligned} \tag{5}$$

Proof Sketch. The posterior conditionals maintain the conditional independence structure proved for the in Theorem 1, i.e. $p(\mathbf{W}_{11},\mathbf{W}_{12},\mathbf{W}_{22\cdot1}|\mathbf{S},\mathbf{T}) = p(\mathbf{W}_{11},\mathbf{W}_{12}|\mathbf{S},\mathbf{T})p(\mathbf{W}_{22\cdot1}|\mathbf{S},\mathbf{T})$. Derivations of the distributions in Eqs. (4) and (5) follow from factorising the posterior and rearranging terms. Relevant calculations are provided in the supplementary document. □

Theorem 2 shows that estimation of the Markov blanket of the p query variables only requires sampling from the posterior conditionals of $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$, which both have an analytical form while remaining independent of $\mathbf{W}_{22}$. Therefore, the amount of parameters in the Markov blanket that need to be estimated, scales linearly with q. This is an improvement over the Bayesian graphical lasso [Wang et al., 2012] approach, where this number grows quadratically with q. Theorem 2 also provides us with the particular distributions to sample from. Next, we demonstrate how this sampling can be done efficiently.

3.1 Efficiency of Sampling from the Posterior

The blockwise Gibbs sampling scheme for estimating the Markov blanket is summarised in Algorithm 1. This sampling scheme consists of iterative resampling of $\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{S},\mathbf{T}$ and of $\mathbf{W}_{11}|\mathbf{W}_{12},\mathbf{S},\mathbf{T}$, according to their definitions in Theorem 2. The estimate of the Markov blanket $\hat{\mathbf{W}}_{12}$ is subsequently computed based on samples drawn from $\mathbf{W}_{12}|\mathbf{W}_{11},\mathbf{S},\mathbf{T}$ following the burn-in period of the sampler.

Algorithm 1: Block Gibbs sampling scheme for the posterior.

Input: Sample covariance matrix $\mathbf{S}$
Output: Markov Blanket estimate $\hat{\mathbf{W}}_{12}$

while *not converged* **do**

$$
\begin{aligned}
T_{ij} &\sim \mathcal{IG}\left(\sqrt{\gamma^2/w_{ij}^2}, \gamma^2\right) \\
vec(\mathbf{W}_{12}^\mathsf{T})|\mathbf{W}_{11}, \mathbf{S} &\sim \mathcal{N}_{pq}\left(vec(-(\mathbf{S}_{22}+\mathbf{I})^{-\mathsf{T}}\mathbf{S}_{12}^\mathsf{T}\mathbf{W}_{11}^\mathsf{T}), \mathbf{C}^{-1}\right) \\
\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S} &\sim \mathcal{MGIG}_{p\times p}\left(-\tfrac{1}{2}(n+p+1), \mathbf{W}_{12}(\mathbf{S}_{22}+\mathbf{I})\mathbf{W}_{21}, \mathbf{S}_{11}+\mathbf{I}\right)
\end{aligned}
$$

return *averaged and thresholded* $\mathbf{W}_{12}$

The distribution of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ is given by Theorem 2(1). The vectorised rows of $\mathbf{W}_{12}|\mathbf{W}_{11}, \mathbf{S}, \mathbf{T}$ follow a joint normal distribution. For $\boldsymbol{v} = vec(\mathbf{S}_{12}^\mathsf{T})$, the distribution further simplifies to

$$
vec(\mathbf{W}_{12}^\mathsf{T})|\mathbf{W}_{11}, \mathbf{S} \sim \mathcal{N}_{pq}\left(-\mathbf{C}^{-1}\boldsymbol{v}, \mathbf{C}^{-1}\right). \quad (6)
$$

The majority of the computational cost incurred in our method arises from sampling from this joint normal distribution. Eq. (6) requires us to invert $\mathbf{C}$, which is of size $pq \times pq$. Note that $\mathbf{C}$ cannot be represented as a covariance tensor of a matrix normal distribution. Therefore, naïve inversion of $\mathbf{C}$ using a standard Cholesky decomposition would cost $\mathcal{O}(p^3q^3)$ operations. Our efficient sampling strategy exploits the structure of this matrix, which is the foundation of Theorem 3.

Theorem 3. *Sampling from the distribution in Theorem 2(1) requires $\mathcal{O}(pq^3)$ operations.*

Proof Sketch. We expand the Kronecker product of matrix $\mathbf{C} \in \mathbb{R}^{pq\times pq}$, which comprises p blocks of size $q \times q$:

$$
\mathbf{C} = \begin{pmatrix}
u_{11}(\mathbf{S}_{22}+\mathbf{I}) + \mathbf{D}_1 & u_{12}(\mathbf{S}_{22}+\mathbf{I}) & \cdots \\
u_{21}(\mathbf{S}_{22}+\mathbf{I}) & \ddots & \\
\vdots & \cdots & u_{pp}(\mathbf{S}_{22}+\mathbf{I}) + \mathbf{D}_p
\end{pmatrix}
$$

where $\mathbf{U} = \mathbf{W}_{11}^{-1}$ is the inverted upper diagonal block. We observe a regular structure within the blocks in $\mathbf{C}$: Matrices $\mathbf{D}_i$ are added to the diagonals blocks only, and the non-diagonal blocks only differ by scalar factors u_{ij}. With a blockwise Cholesky factorisation, the inversion requires only pq^3 operations. Since the Cholesky decomposition of the blocks also only differs by a factor, we can store its intermediate result. □

Remark If there are further memory constraints, distributed versions of the Cholesky decomposition should be considered to enhance performance.

Theorem 2(2) states that $\mathbf{W}_{11}|\mathbf{W}_{12}, \mathbf{S}, \mathbf{T}$ follows the MGIG distribution. In order to sample from this distribution, we make use of a result by Bernadac [1995]. It introduces a representation of an MGIG-distributed random variable as a limit of a random continued fraction of Wishart-distributed random variables. The interested reader should refer to Bernadac [1995]; Koudou et al. [2014]; Letac [2000] for the details. Drawing samples from the MGIG thus reduces to iterated sampling from the Wishart distribution. In practice, we observe the convergence of the random continued fraction within few iterations. The complexity of sampling from the distribution derived in Theorem 2(2) does not depend on q.

4 EXTENSION WITH GAUSSIAN COPULA

We extend the model for non-Gaussian and mixed continuous/discrete data by embedding it within a copula construction. Copulas describe the dependency in a r-dimensional joint distribution $F(Y_1, \ldots, Y_r)$ and represent an invariance class with respect to the marginal cumulative distribution functions (cdf) F_i. In our model, $r = p + q$. For continuous cdfs, Sklar's theorem [Sklar, 1959] guarantees the existence and uniqueness of a copula C, such that $F(Y_1, \ldots, Y_r) = C\left(F_1(Y_1), \ldots, F_r(Y_r)\right)$. For discrete cdfs, this leads to an identifiability problem [Genest and Neslehova, 2007], such that established methods on empirical marginals [Liu et al., 2009] cannot be used anymore, but a valid copula can still be constructed [Genest and Neslehova, 2007]. For our purpose, we follow the semi-parametric approach by Hoff [2007] and restrict our model to the parametric Gaussian copula, but we do not restrict the data to be Gaussian and treat them in a non-parametric fashion. The Gaussian copula inherently implies latent variables $X_i = \Phi^{-1}(F_i(Y_i))$. Our model under consideration is

$$
(X_1, \ldots, X_r)^\mathsf{T} \sim \mathcal{N}_r\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right), \quad Y_i = F_i^{-1}\left(\Phi(X_i)\right) \quad (7)
$$

where F_i^{-1} denotes the ith generalised inverse of continuous or discrete cdfs, $\mathbf{X}$ are the latent variables, and $\mathbf{Y}$ are the observations.

Following Hoff [2007], inference in the latent vari-

ables uses the non-decreasing property of discrete cdfs for transforming the observed variables to the latent space. This guarantees that for observations $y_{ik} < y_{il}$ we also have $x_{ik} < x_{il}$, and more generally, $\mathbf{X}$ must lie in the set

$$\mathcal{D} = \{\mathbf{X} \in \mathbb{R}^{r \times n} : \max(x_{ik} : y_{ik} < y_{ij}) < x_{i,j} < \min(x_{ik} : y_{ij} < y_{ik})\}$$

The data likelihood can then be written as

$$p(\mathbf{Y}|\mathbf{\Sigma}, F_1, \ldots, F_r) = p(\mathbf{X} \in \mathcal{D}, \mathbf{Y}|\mathbf{\Sigma}, F_1, \ldots, F_r) = p(\mathbf{X} \in \mathcal{D}|\mathbf{\Sigma}) p(\mathbf{Y}|\mathbf{X} \in \mathcal{D}, \mathbf{\Sigma}, F_1, \ldots, F_r)$$

and estimation of $\mathbf{\Sigma}$ is performed on maximising the sufficient statistics $p(\mathbf{X} \in \mathcal{D}|\mathbf{\Sigma})$ only, thus treating the marginals F_i as nuisance parameters. Bayesian inference for $\mathbf{\Sigma}$ is achieved by a Markov chain having stationary distribution at the posterior $p(\mathbf{\Sigma}|\mathbf{X} \in \mathcal{D}) \propto p(\mathbf{\Sigma}) p(\mathbf{X} \in \mathcal{D}|\mathbf{\Sigma})$, where a inverse-Wishart prior $p(\mathbf{\Sigma})$ is used. Posterior inference can be achieved with a Gibbs sampler, which draws alternately between $\mathbf{X}|\mathbf{\Sigma}, \mathbf{Y}$ and $\mathbf{\Sigma}|\mathbf{X}$. This sampler extends Alg. 1 with an additional outer loop for inferring the latent variables. The Markov blanket is then iteratively estimated on these variables. The sampling scheme easily accommodates for missing values, when omitting conditioning on the set $\mathcal{D}$.

The presented framework is very useful in practice, since the invariance class of copulas extend the model to non-Gaussian data. With the additional stochastic transformation to the latent space, we can use discrete variables and allow missing values. In real world applications, it becomes apparent that this is a very valuable extension.

5 EXPERIMENTS

5.1 Artificial Data

As a first experiment, we attempt to highlight the differences in inference between the Bayesian Markov blanket (BMB) and Bayesian Graphical Lasso (BGL) procedures. We construct an artificial network with 100 variables, where our interest is confined to only the Markov blanket between $p = 10$ query variables and the $q = 90$ remaining variables. In order to create networks with a "small-world" flavour containing *hubs*, i.e. nodes with very high degree, the connectivity structure of the inverse covariance matrix $\mathbf{W}$ is generated by a beta-binomial model. Edge weights are sampled uniformly from the interval $[0.3, 1]$, and edge signs are randomly flipped. Finally, positive definiteness is guaranteed by adding a suitable constant (related to the smallest eigenvalue) to the diagonal. This process produces a sparse network structure where the majority of edges are connected to only a few single nodes. Note that many real-world networks exhibit such small-world properties.

Next, we draw $n = 1\,000$ independent samples from a zero-mean normal distribution with covariance matrix $\mathbf{W}^{-1}$ and compute the sample covariance $\mathbf{S}$. Fig. 2 depicts a true Markov blanket and its reconstruction by BGL and BMB using the same sparsity parameter $\lambda = 200$. Both methods were run side-by-side for 700 MCMC samples after an initial burn-in phase of 300 samples. From the sampled networks, a representative network structure is constructed by thresholding based on a 85% credibility interval. We repeat the above procedure to obtain a total of 100 datasets. The quality of reconstructed networks is measured in terms of f-score (harmonic mean of precision and recall) between the true and inferred Markov blanket. When computing precision and recall, inferred edges with edge weights having the wrong sign are counted as missing. Both models share the same sparsity parameter λ, which in this experiment was selected such that for BMB recall and precision have roughly the same value. The results are depicted as box plots in Fig. 3, from which we conclude that there are indeed substantial differences in both models. In particular, BGL has the tendency to introduce many unnecessary edges in comparison to BMB. As a result, BGL achieves high recall and low f-score. Since both methods are based on the same likelihood model and (almost) the same prior, the observed differences can only be attributed to differences in the inference procedure: BGL infers a network by iterating over *all* variables and their neighbourhood systems, whereas BMB only estimates the elements in $\mathbf{W}_{11}$ and $\mathbf{W}_{12}$.

To further study the influence of the different Gibbs sampling strategies, we examine tracer plots and autocorrelations of individual variables in Fig. 4. In almost all cases, BGL shows significantly higher autocorrelation and poor convergence. In contrast, Markov chains in the BMB sampler seems to mix much better, typically leading to posteriors with smaller bias and variance. While only one example is shown in the figure, similar results can be seen for basically all variables in the network. Further, we experience a substantial decrease in run-time, even for these relatively small networks: computing 1 000 MCMC samples for BMB finished on average after 100 seconds, while BGL typically consumed around 370 seconds. Since BGL requires an additional sampling loop over *all* variables, datasets with large $\mathbf{S}_{22}$ quickly become problematic for BGL. We further explore these differences in the next section for a large real-world application.

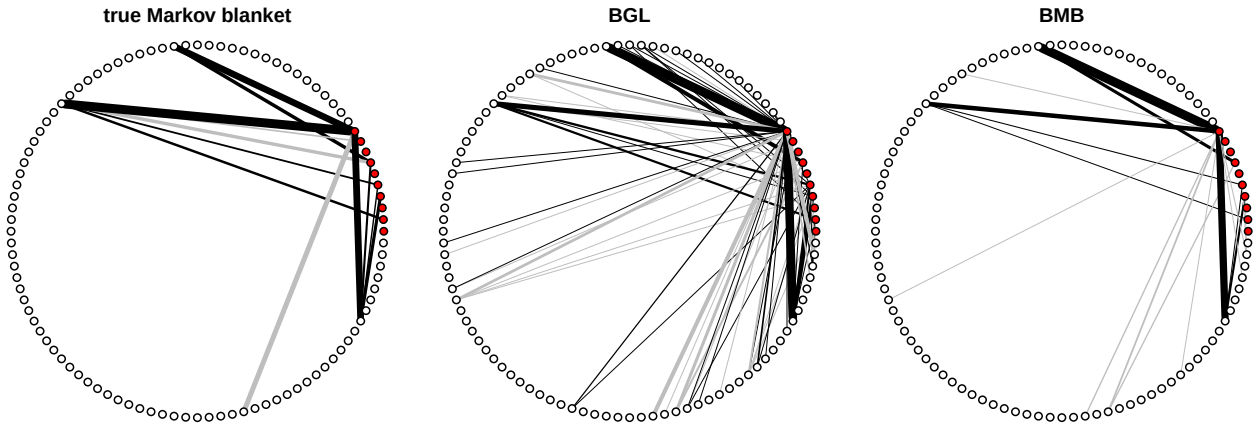

Figure 2: One exemplary Markov blanket ($p = 10$, $q = 90$) and its reconstruction by BGL and BMB. Note that the graphs *only* display edges between p query and q remaining variables. Red nodes represent query variables, white nodes represent all other variables. Black and grey edges correspond to positive and negative edge signs, respectively.

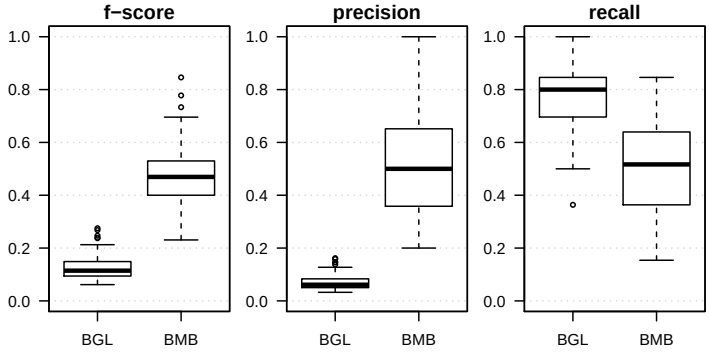

Figure 3: Performance of inferred Markov blankets from 100 datasets.

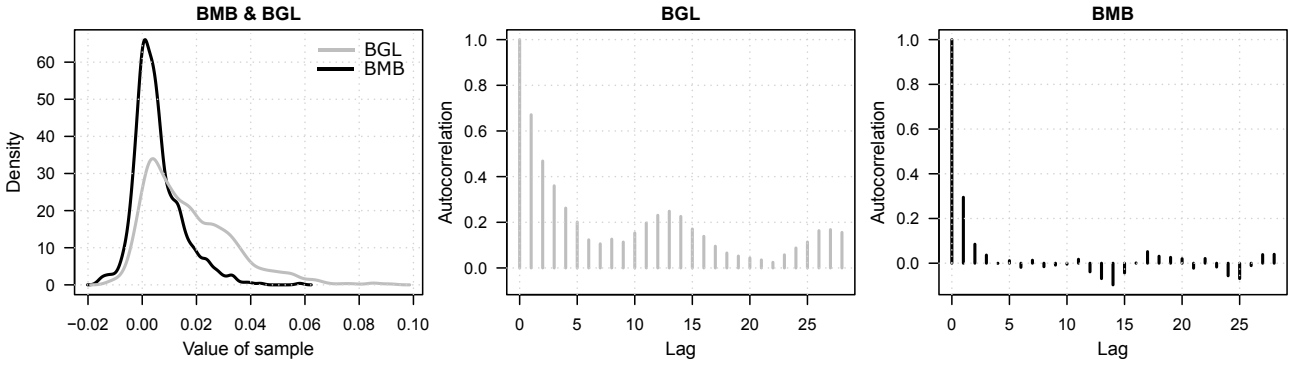

Figure 4: Density and auto-correlation of the Markov chain for a single variable in the Markov blanket. Gray refers to BGL, black to BMB.

5.2 Real Data

To demonstrate the practical significance of Markov blanket estimation, we turn to the analysis of *colorectal cancer*, which in 2012 ranked among the three most common types of cancer globally [Stewart and Wild, 2014]. The data set introduced in Sheffer et al. [2009] is publicly available and contains gene expression measurements from biopsies of $n = 260$ cancer patients. A separate table captures discrete/categorical clinical traits such as sex, age or pathological staging/grading. In this context, one particularly interesting research question is to identify connections between the p (macroscopic) clinical descriptors and the q (molecular) gene expression measurements.

Among the 13 400 genes contained in the dataset, we focus on a specific subset, the so-called "*Pathways in cancer*" as defined in the KEGG database[2]. This particular subset comprises a general class of genes which are known to be involved in various biological processes linked to cancer. For this experiment, we have $q = 312$ candidate genes and $p = 7$ query variables. These are the age and sex of the patient as well as the *TNM* classification, cancer group stage (GS) and mutation of the tumor suppressor protein *p53*. Since the observations have mixed continuous/discrete data types with missing values, the Markov blanket estimation is extended by a semi-parametric Gaussian copula framework [Hoff, 2007]. Based on this, we calculate 5 000 MCMC samples, which finally leads to the Markov blanket in Fig. 5.

The resulting network structure confirms some well-known properties like the confounding effect of the age and sex variables, both of which (correctly) link to a large number of genes. For example, *FGF21* exhibits significant differences in male and female subjects [Bisgaard et al., 2014], and *CTNNA1* shares connections to survival time in men [Ropponen et al., 1999]. Similarly, *mTOR*, the *mechanistic target of rapamycin*, not only represents a key element for cell signaling that triggers a cascade of immune-related pathways, but its function also depends heavily on a subject's age [Johnson et al., 2013]. Despite these age- and sex-related observations being non-trivial, they are not of primary interest, which is why the remaining variables carry more practical insights from a clinical point view. Further, we are able to identify a very interesting network structure around the variable *tumor size* T: almost all direct neighbours control either cell growth (*EGLN1* [Erez et al., 2003], *RELA* [Yu et al., 2004], *HGF* [Date et al., 1998; Renzo et al., 1995] and others) or cell death (*BCL2*, *FADD*). Cancer typically affects the balance between these two fundamental processes and their deregulation eventually leads to tumor development. A second subgraph concerns variable N, the degree of spread to regional lymph nodes, which is expressed in 4 levels $N0$ to $N3$. Here, all genes in the neighbourhood correspond to the lymphatic system and its direct responses to malignant cell growth, which was confirmed for *FGF9* [Deng et al., 2013], *MDM2* [Fridman et al., 2003; Leitea et al., 2001] and *TRAF4* [Camilleri-Broet et al., 2007] among others. Finally, the following two clinical variables appear to be conditionally independent from genes, yet they may internally depend on other clinical variables (i.e., outside of the Markov blanket): binary M (presence of metastasis in distant organs) and discrete GS (group stage of cancer). Interestingly, the latter is only a summary function of T, N and M, hence internal links to the aforementioned variables are very likely.

Despite the study's focus on colorectal cancer and specifics of the intestinal system, the inferred Markov blanket is able to explain rather general properties in accordance with findings in the medical literature. Altogether, this nicely illustrates how the Gaussian copula framework complements the Bayesian Markov blanket estimation – especially pertaining to the clinical domain with mixed observations and missing values.

In contrast to our approach, the high dimensionality of this dataset imposes severe problems for BGL. For BMB, 5 000 Gibbs sweeps could be computed in 2 hours, and MCMC diagnosis did not show any severe convergence problems. For BGL, however, the same number of iterations already took 122 hours (≈ 5 days), and we observed similar (and sometimes severe) mixing problems as described in the previous section.

5.3 A Note About The Graphical Lasso

A natural question that arises is how the BMB solution presented here compares to existing frequentist techniques, particularly the classical graphical lasso due to Friedman et al. [2008]. The BMB uses the same likelihood as the graphical lasso. As a result, comparing both techniques reduces to comparing Bayesian inference with maximum likelihood inference. Evidently, such a comparison reveals that the BMB provides us with a posterior distribution that expresses our confidence in a solution, while the graphical lasso only returns a point estimate. It should also be noted that BMB and the graphical lasso are virtually identical if a highly peaked prior is used.

[2]Kyoto Encyclopedia of Genes and Genomes, `http://www.genome.jp/kegg/pathway.html`

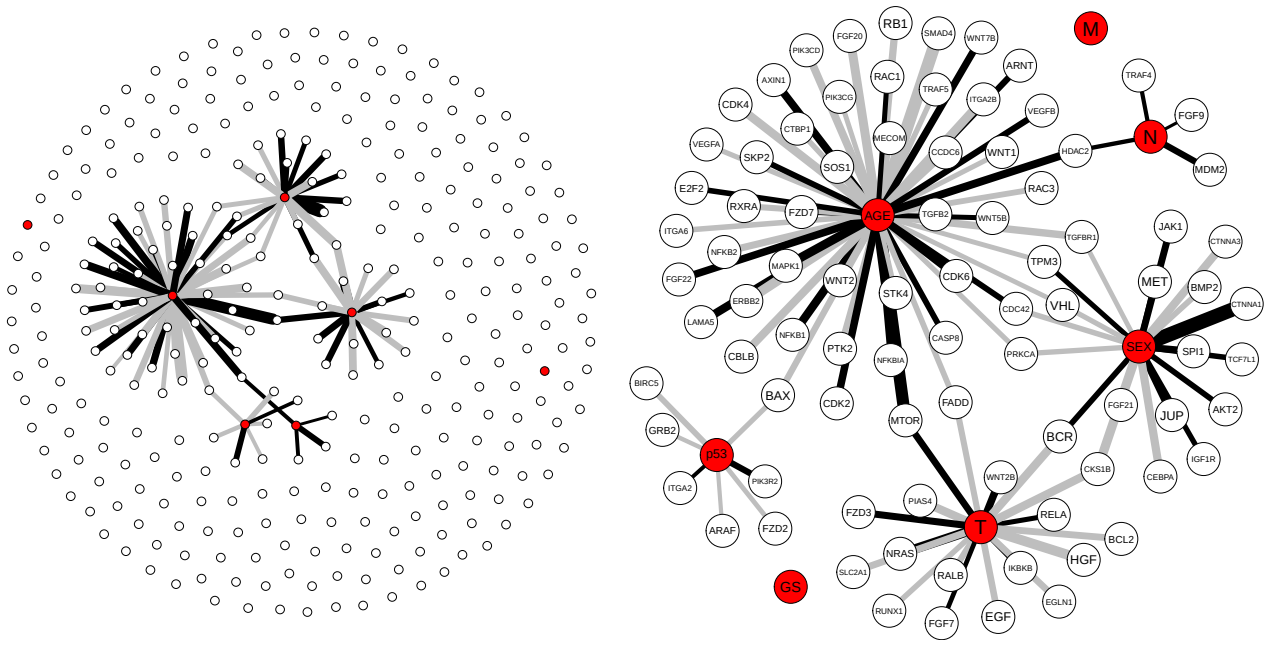

Figure 5: Sparse Markov blanket between $p = 7$ clinical features (red nodes) and $q = 312$ genes in colorectal cancer [Sheffer et al., 2009]. Overview of all variables/nodes (left) and enlarged, fully labeled subgraph (right).

6 CONCLUSION

We have presented a Bayesian perspective for estimating the Markov blanket of a set of query nodes in an undirected network. In our experience, it is often the case that we estimate a full network but interpret only part of it. This is especially true in a context where portions of our data are qualitatively different. Here, we would be more interested in establishing the links between these portions, rather than examining the links within the portions themselves. Markov blanket estimation is hence an interesting and relevant sub-problem of network estimation, particularly in high dimensional settings. Existing methods such as the Bayesian graphical lasso iterate through the individual variables to estimate an entire network. While there are several situations in which inference of the entire network is required, there are also cases in which we are only interested in the neighbourhood of a small subset of query variables; for these instances, iterating through all the variables is unnecessary.

In this paper, we explored the blockwise factorisation of the Wishart likelihood in combination with a suitable choice of prior. Our primary contribution in Theorem 2 shows that the resulting posterior distribution of the Markov blanket of a set of query nodes has an analytic form, and is independent of a large portion of the network. The analytic form allows us to explore potentially large neighbourhoods where the Bayesian graphical lasso reaches its limits. We also demonstrated that sampling from the posterior of the Markov blanket is more efficient than the Bayesian graphical lasso. Moreover, we observed fast convergence and superior mixing properties of the Markov chain. We attribute this to the improved flexibility of our sampling strategy.

Including a copula construct in the model further enhances its real world applicability, where mixed data and missing values are prevalent. A particular application in a medical setting is the colorectal example we considered in Section 5.2. Using this approach allowed us to make interesting observations about the interactions between various clinical and genetic factors. Such insights could ultimately contribute to a better understanding of the disease.

In the spirit of research reproducibility, we provide source code in the supplement.

Acknowledgement This work was partially supported by the Swiss National Science Foundation, projects 200021_146178, CR32I2_159682, 51MRP0_158328.

References

Evelyne Bernadac. Random continued fractions and inverse gaussian distribution on a symmetric cone. *Journal of Theoretical Probability*, 8(2):221–259, 1995.

A. Bisgaard, K. Sørensen, T. H. Johannsen, J. W. Helge, A.-M. Andersson, and A. Juul. Significant Gender Difference in Serum Levels of Fibroblast Growth Factor 21 in Danish Children and Adolescents. *International Journal of Pediatric Endocrinology*, 2014(1):7, 2014.

Ronald W Butler. Generalized inverse gaussian distributions and their wishart connections. *Scandinavian journal of statistics*, 25(1):69–75, 1998.

S. Camilleri-Broet, I. Cremer, B. Marmey, E. Comperat, F. Viguie, J. Audouin, M.-C. Rio, W.-H. Fridman, C. Sautes-Fridman, and C. H. Regnier. TRAF4 Overexpression is a Common Characteristic of Human Carcinomas. *Oncogene*, 26(1):142–147, 2007.

K. Date, K. Matsumoto, K. Kuba, H. Shimura, M. Tanaka, and T. Nakamura. Inhibition of Tumor Growth and Invasion by a Four-Kringle Antagonist (HGF/NK4) for Hepatocyte Growth Factor. *Oncogene*, 17(23):3045–3054, 1998.

M. Deng, H. Tang, X. Lu, M. Liu, X. Lu, Y. Gu, J. Liu, and Z. He. miR-26a suppresses tumor growth and metastasis by targeting FGF9 in gastric cancer. *PloS one*, 8(8): e72662, 2013.

N. Erez, M. Milyavsky, R. Eilam, I. Shats, N. Goldfinger, and V. Rotter. Expression of Prolyl-Hydroxylase-1 (PHD1/EGLN2) suppresses Hypoxia Inducible Factor-1α Activation and Inhibits Tumor Growth. *Cancer Research*, 63(24):8777–8783, 2003.

J. S. Fridman, E. Hernando, M. T. Hemann, E. de Stanchina, C. Cordon-Cardo, and S. W. Lowe. Tumor Promotion by MDM2 Splice Variants Unable to Bind p53. *Cancer Research*, 63(18):5703–5706, 2003.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

C. Genest and J. Neslehova. A primer on copulas for count data. *Astin Bulletin*, 37(2):475, 2007.

Arjun K Gupta and Daya K Nagar. *Matrix variate distributions*, volume 104. CRC Press, 1999.

Peter D Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.

S. C. Johnson, P. S. Rabinovitch, and M. Kaeberlein. mTOR is a Key Modulator of Ageing and Age-Related Disease. *Nature*, 493(7432):338–345, 2013.

A. Efoévi Koudou, C. Ley, et al. Characterizations of GIG laws: A survey. *Probability Surveys*, 11:161–176, 2014.

K. R. M. Leitea, M. F. Franco, M. Srougi, L. J. Nesrallah, A. Nesrallah, R. G. Bevilacqua, E. Darini, C. M. Carvalho, M. I. Meirelles, I. Santana, and L. H. Camara-Lopes. Abnormal Expression of MDM2 in Prostate Carcinoma. *Modern Pathology*, 14(5):428–436, 2001.

Gerard Letac. Symmetric cones as gelfand pairs: probabilistic applications. *Contemporary Mathematics*, 261: 109–120, 2000.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462, 2006.

M. F. Di Renzo, M. Olivero, A. Giacomini, H. Porte, E. Chastre, L. Mirossay, B. Nordlinger, S. Bretti, S. Bottardi, and S. Giordano. Overexpression and Amplification of the Met/HGF Receptor Gene During the Progression of Colorectal Cancer. *Clinical Cancer Research*, 1(2):147–154, 1995.

K. M. Ropponen et al. Reduced Expression of alpha Catenin is Associated with Poor Prognosis in Colorectal Carcinoma. *Journal of Clinical Pathology*, 52(1):10–16, 1999.

M. Sheffer, M. D. Bacolod, O. Zuk, S. F. Giardina, H. Pincas, F. Barany, P. B. Paty, W. L. Gerald, D. A. Notterman, and E. Domany. Association of Survival and Disease Progression with Chromosomal Instability: A Genomic Exploration of Colorectal Cancer. In *Proceedings of the National Academy of Sciences*, pages 7131–7136, 2009.

M Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

B. W. Stewart and C. P. Wild. *World Cancer Report 2014*. IARC Press, 2014.

Hao Wang et al. Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4): 867–886, 2012.

H.-G. Yu, X. Zhong, Y.-N. Yang, H.-S. Luo, J.-P. Yu, J. J. Meier, H. Schrader, A. Bastian, W. E. Schmidt, and F. Schmitz. Increased Expression of Nuclear factor-κB/RelA is Correlated With Tumor Angiogenesis in Human Colorectal Cancer. *International Journal of Colorectal Disease*, 19(1):18–22, 2004.